Capstone #1 Final Report

Problem Statement

European soccer clubs spend enormous amounts of money on new players every year. I want to see if clubs that spend more on new players typically see an improvement in the results they achieve on the field. Soccer club executives can use this to help guide how much they choose to spend on players, while scouts will be able to use this to inform them on which leagues are producing players that have a significant impact on results. The data will come from the following datasets:

https://www.kaggle.com/hugomathien/soccer
https://github.com/ewenme/transfers/tree/master/data

The first dataset consists of match details of every game played in the top division of the seven European soccer leagues of interest (English, Spanish, German, French, Italian, Dutch, and Portuguese), including the teams involved, the goals scored by each team. The second dataset contains every player bought and sold by each team in these leagues of interest, along with attributes about each player, including age, position, and the other team involved in the sale.

Data Wrangling

The match data was held in a SQLite database. I used the sqlite3 and pandas package to read the necessary tables in as pandas dataframes. The table containing match data used only ID numbers for teams and leagues, so I needed to also read in the tables containing the corresponding team and league names. I created dictionaries consisting of team ID : team name as well as league ID : league name, as key, value pairs in order to add the correct team and league name to each row in the match dataset. I also added three boolean columns, home_win, away_win, and draw. One of the three columns in each row would read True depending on the result of the match. This would be used to total a given team's record at the end of a season.

The only column in this dataset which contained an inconsistent data type was the year column. In soccer, seasons span across two years and are typically referenced by both years. For example, the 2008/2009 season references the season that starts in late 2008 and spans until mid-2009. Because of this, this column consisted of strings containing both years ("2008/2009", "2009/2010", etc.). To fix this, I changed the year to an integer of the first four characters, so each season would be referenced by the year it started (2008 for 2008/2009 season). This would allow for easier and cleaner plotting by season.

In order to assess a team’s performance on a yearly basis, I created a function that outputs an end-of-season league table to display what place each team finished in a given year and given league. These league tables also display each team’s wins, losses, draws, goals scored, goals conceded, win percentage, and position value. I defined position value as 100-5*(league_position-1), so it is 100 for a first place finish, and it decreases by 5 with each place. I created this metric in order to assign a higher value for finishing higher in the table, which seems more intuitive when plotting data than using the actual league position, which is a lower number the higher a team finishes.

The transfer data is contained in a collection of CSV files grouped by year and league. I looped through each year, read in the CSV of each league’s transfers, appended each dataframe to a list, and then concatenated the list together to make a single dataframe containing all of the transfer data for the time period.

To make the transfer data more manageable, I first split the dataset into two subsets, whether the transfer was a purchase or sale. There were too many specific values for player position, so I replaced each value with Forward, Midfield, or Defender where appropriate. I also added an “age_range” column which had the value 15-22, 23-29, or 30+ depending on age. This would allow for future grouping to see how teams spent differently on different positions and age groups.

In order to match a team’s performance data to the corresponding transfer data, I added an ID column to the transfer data and populated it with the ID from the match dataset. This only worked, however, with team names that were exact matches across the two datasets, as some team names differed slightly from the different sources. For example, “Arsenal FC” is sometimes referred to simply as “Arsenal”, which Python sees as different strings despite referencing the same team. In order to populate the ID column for teams without perfect matches, I used the partial_ratio function from the fuzzywuzzy.fuzz package to obtain the string comparison ratio and populated the ID number column if the strings met the necessary threshold.

Then, in order to properly prepare my data for modeling, I compiled all of the necessary information into a single dataframe that contained the amount each team spent per year, along with a breakdown on how much each team spent per age group and position group. I also included the team’s performance statistics per year, like the position value, win percentage, goals scored, and conceded.

Exploratory Analysis

In European soccer, teams that are in the top division have much more money to spend on players due to television deals, greater ticket and merchandise sales, and wealthier owners who, in most cases, invest more of their own money back into the team. Teams that are able to stay in the top division in the team's respective country will have a consistently higher budget than those who are relegated to lower divisions, even for a short time. For this reason, the clubs of interest for my analysis were teams that remained in the top division for all 10 years between 2007 and 2016.

Figure 1

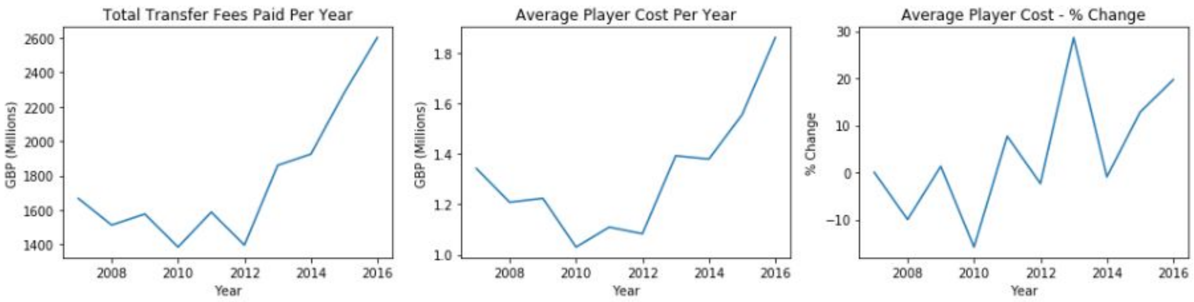


Figure 1 shows the total amount spent by all of the clubs of interest per year, average cost of player per year, and the average player cost percent change per year. Teams in 2013 started spending significantly more than in previous years, as shown by the near 30% increase in average player cost. There were also sharp increases in total money spent from 2013-2016, despite the slight decrease in average player cost in 2014, though the average player cost in 2014 was still higher than the previous high for the period in 2007.

Figure 2

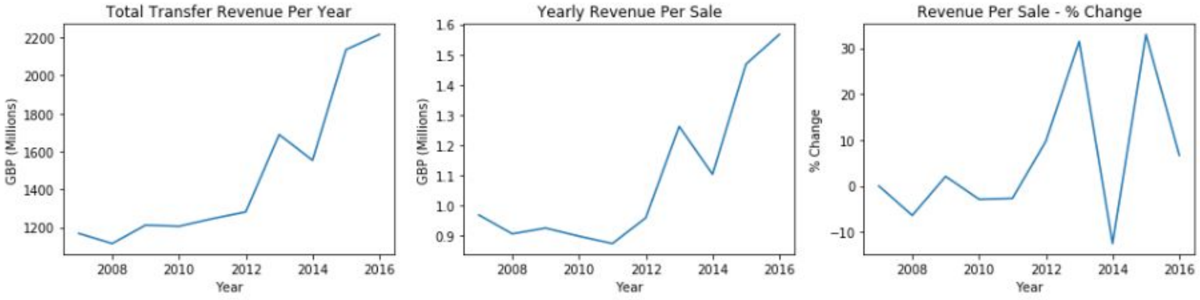


Figure 2 shows the total transfer revenue for all of the clubs of interest per year, average revenue per sale, and revenue per sale percent change. Perhaps one reason top teams are spending much more on players is that they have been able to generate more money by selling their own players.

The revenue for top teams follows a similar trend to the money they are spending, with the sharp increases in 2013 and 2015, in addition to another slight increase in 2016. Another reason for the increases in this time period could be the influx of outside money into transfer business. Many clubs in recent years have experienced takeovers by extremely wealthy ownership groups who put a lot of their own money into their club. This not only allowed for a much higher budget for these few suddenly richer clubs, but it also forced rival clubs to spend more than they are used to in order to remain competitive.

Figure 3

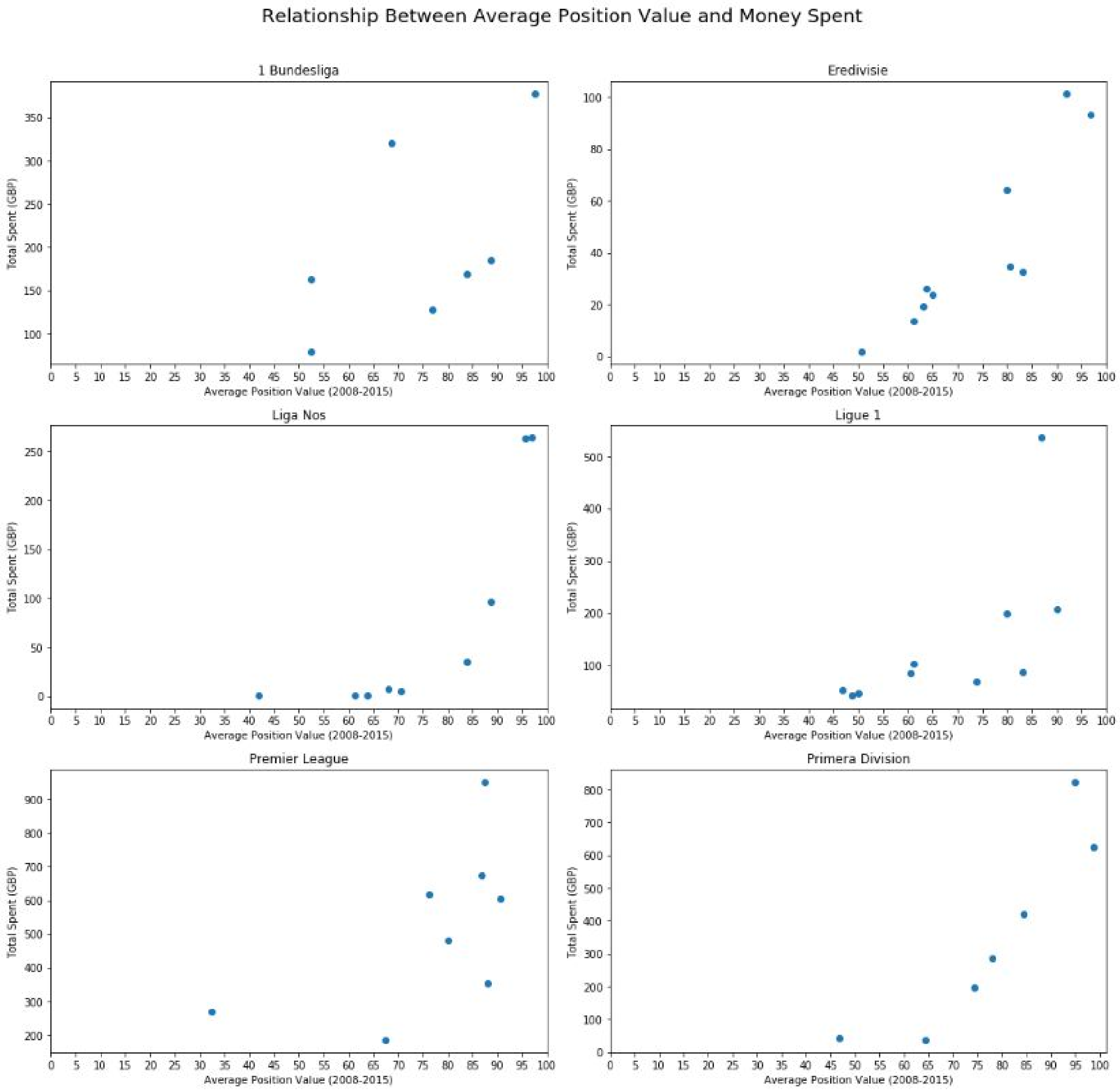


Figure 3 (continued)

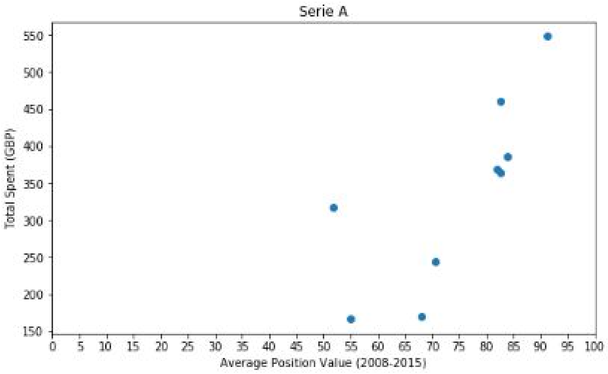


Figure 3 shows the relationship between the average position value for clubs of interest during 2008-2015 and the total amount spent. I decided to make a separate plot per league because some league’s have a drastically different budget compared to others. For example, the top spenders in the Eredivisie (Dutch League), spent around 100 million pounds, while the top spenders in the Premier League (English League) and Primera Division (Spanish League) are closer to 1 billion pounds.

Figure 4

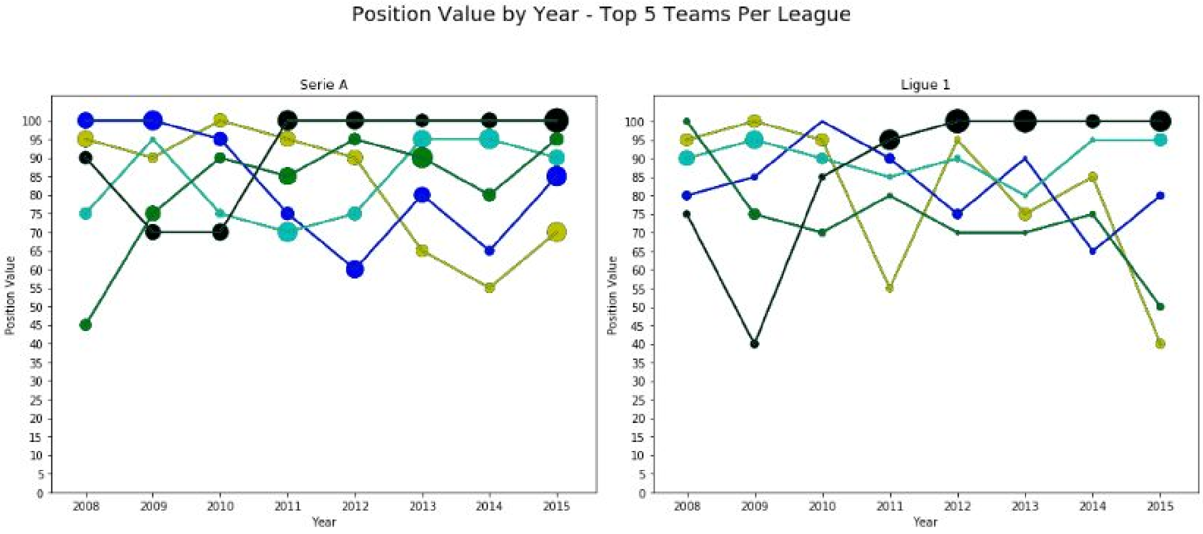
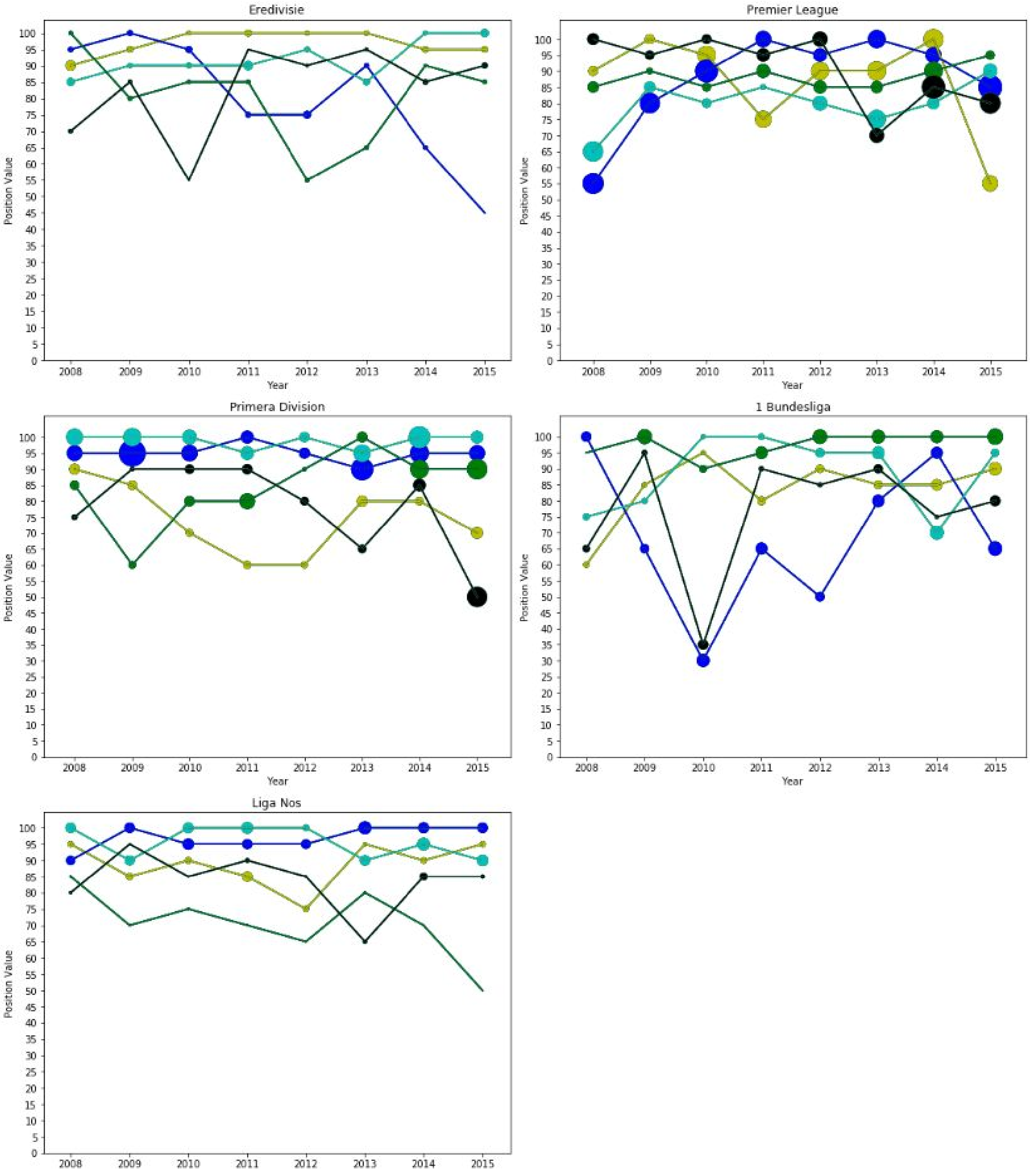


Figure 4 shows the change in position value by year of the teams in each league with the top 5 position values. The dot size in a given year is the team's money spent. In the smaller leagues such as the Eredivisie (Dutch League), Liga Nos (Portuguese League), and Ligue 1 (French League), the teams with higher position values tend to have spent more money, whereas in the league's known to be more competitive, such as the Premier League (England) and Primera Division (Spain), teams that spend the most seem to win less often.

Figure 4 (continued)


Eredivisie

Premier League

Primera Division

1 Bundesliga

Liga Nos

Even though the league winners shown in Figure 4 are not always the year’s top spender, higher spending looks to at least put teams in contention for a league title on a consistent basis.

Statistical Data Analysis

I first tested the correlation between all variables with a correlation matrix, as well as two OLS linear regression models with position value as the dependent variable. I found that there is not a strong direct correlation between the position value and any of the spending categories, with R-squared values for these models at around 0.1. There is, however, a strong correlation between position value and the performance categories, such as goals scored, goals against, and goal difference. This is to be expected, as it clearly follows that if a team scores more and concedes less, they will win more games and finish higher in the league table.

Because of this, I decided to look at the relationship between the performance categories and spending categories. Goal difference is highly correlated to position value (correlation coefficient of 0.84), so I generated OLS models with goal difference as the dependent variable and the spending categories as the independent variables. The R-squared values for these models were slightly higher than the previous set of models at almost 0.2, but these would still not suggest a strong correlation between spending and results. Despite the low R-squared values for these models, with the exception of the 30+ category, the p-values for each variable were very low, so I would reject the null hypothesis that these variables have no effect on the dependent variable.

One reason for the low R-squared values in these models could be that it takes more than a year for a team to see an improvement in results after spending a significant amount of money, so I created two additional columns in the dataframe. The first is a two year total amount spent (null for the first year of data for each team), and the second is a three year total amount spent (null for the first two years of data for each team. I generated two OLS models, both with goal difference as the dependent variable, one with two year total as the independent variable, and one with three year total as the independent variable. The R-squared values for the two and three year models were 0.23 and 0.27, respectively, so while it did slightly increase from the previous models, it was not by a significant amount.

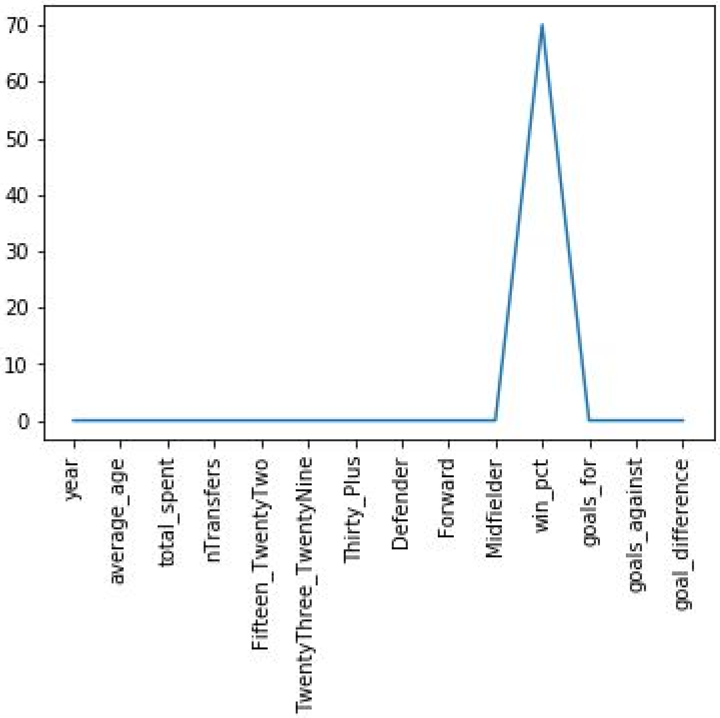
The last method I tried to account for the time it can take for money spent to affect results was to take the average over the time period of data collected. I grouped the original dataframe by team and took the average of each category to create a dataframe with a single row per team and columns containing the average values of all of the original categories. I generated an OLS model with average goal difference as the dependent variable and all of the spending categories as well as average age as the independent variables. With an R-squared of 0.52, this model explains significantly more of the variance in goal difference than any of the previous models, while the most significant variables look to be midfielders, players between 15 and 22, and the average age of players bought.

Many top teams are known to have extraordinary youth academies where they develop young players and integrate them into the full team at no cost to the club, which could account for why some teams have achieved good results without spending much money. Another reason is that many teams take advantage of free transfers, which are players who are out of contract and do not have a parent club that they need to pay in order to sign them to a contract.

In-Depth Analysis

I applied two different machine learning techniques to my data. I first wanted to test for the most important independent variable I had collected, so I chose to use a Lasso regression, so as to shrink all of the unimportant variables to zero. I first split my data into train and test subsets with the sklearn.model_selection tool train_test_split. Then, I made a Lasso model, fit the training set to my data, and checked the coefficients and score for the test set. The only variable that was not shrunk to zero in this model was win percentage, which had a Lasso coefficient of 70.07 (Figure 5). The test set had a score of 0.598, so about 60% of the variance in this model can be explained by win percentage. The root mean squared error was about 13.5, which is high, leading me to believe it is more than just the win percentage that can explain a team's position value.

Figure 5 - Lasso Regression Coefficients



I also implemented random forests in my data, so as to see the importances of my features, but not shrink all but one of them to zero. I created a random forest regressor with the sklearn.ensemble package and fit the model to the training set. Obtaining the score on the test set yielded a value of 0.857, which is significantly higher than the Lasso regression. The root mean squared error was also lower at 8.09. The most important feature was still win percentage at 0.64, but goal difference (goals scored minus goals allowed) could also be considered important at 0.24. All of the teams transfer data showed low scores on these models, leading me to conclude that a team's game performances are generally not affected by their spending.